

Context Tricks for Cheap Semantic Segmentation

Thanapong Intharrah and Gabriel J. Brostow
 University College London
 Gower Street, London, United Kingdom
 [t.intharrah,g.brostow]@cs.ucl.ac.uk

Abstract

Accurate semantic labeling of image pixels is difficult because intra-class variability is often greater than inter-class variability. In turn, fast semantic segmentation is hard because accurate models are usually too complicated to also run quickly at test-time. Our experience with building and running semantic segmentation systems has also shown a reasonably obvious bottleneck on model complexity, imposed by small training datasets. We therefore propose two simple complementary strategies that leverage context to give better semantic segmentation, while scaling up or down to train on different-sized datasets.

As easy modifications for existing semantic segmentation algorithms, we introduce Decorrelated Semantic Texton Forests, and the Context Sensitive Image Level Prior. The proposed modifications are tested using a Semantic Texton Forest (STF) system, and the modifications are validated on two standard benchmark datasets, MSRC-21 and PascalVOC-2010. In Python based comparisons, our system is insignificantly slower than STF at test-time, yet produces superior semantic segmentations overall, with just push-button training.

1. Introduction

For many applications, such as navigation or robot-interaction, semantic segmentation of images needs to be both *accurate* and *fast* to be worthwhile. The environment can change more or less abruptly, but typically, many frames will have combinations of the same frequently co-occurring classes. We leverage this persistence of context to improve pixel classification accuracy, given finite quantities of training data.

We build on the successful Semantic Texton Forest (STF) [22] approach, and enhance it through two main contributions. First, the Decorrelated Semantic Texton Forest (DSTF) is proposed as a variant to the STF that essentially preserves the original’s efficiency. The DSTF uses hierarchical clustering to decorrelate classes that have confus-

ingly similar appearance for an STF. We further improve accuracy by introducing a Context Sensitive Image Level Prior (Context Sensitive ILP). Training this multi-label prior to account for the co-occurrence of classes proves to be very helpful and substantially better than the more typical multi-class training of ILPs.

2. Related Work

Many semantic segmentation algorithms require carefully tuned models and/or fully connected Conditional Random Fields (CRFs) to produce accurate per-pixel labelings. Here, we review the most relevant such methods, as well as algorithms similar to ours that are close to state-of-the-art, but sacrifice some accuracy for improved test-time efficiency.

General Algorithms

We base our approach on the basic STF because we wish to leverage its low computational complexity, allowing for very fast implementations if needed. Shotton *et al.* introduced STF as a component of their Bag of Semantic Textons (BoST) [22] model. BoST is one of the earliest algorithms to still be competitive in semantic segmentation challenges, appearing on the leaderboards of many semantic segmentation papers [1, 11, 24]. BoST still outperforms other state of the art algorithms in some categories, as shown in Table 4, despite running in real-time. The two main components of the BoST model were i) use of the newly introduced STF, and ii) application of the Textonboost [23] approach for encoding local context information, generating the BoST for each patch. We henceforth refer to the first component as STF, and to their combined approach as BoST. The STF is trained by growing extremely randomized trees with raw pixel image patches as features. Leaf nodes store the class distributions of the image patches that reached that node. Although the whole STF process is considered very efficient at inference time, the STF by itself produces fairly low quality results, because raw pixel patches are often not expressive enough to be discriminative between classes. BoST improves the results dramatically, but with some computa-

tional overhead. Our proposed system modifies the STF by adding only a little overhead, but achieves significantly better performance compared with BoST [22].

One approach known for using an image level prior is [7]. Their overall system has a chain of stages: i) extracting patches and their low-level features, ii) constructing high-level features (Fisher vectors), iii) training (predicting, in test time) a class scoring unit using the high-level features, iv) assigning scores to a pixel and propagating scores to oversegmented regions, and finally v) integrating with their image level prior to refine their labels. Comparing to us, we skip (i) and (ii) which are bottle neck of the algorithm and use DSTF which very efficient because no feature extraction is required; their image level prior does not exploit co-occurrence statistics but model only the presence of each class individually. By combining three simple components: local appearance scoring, context sensitive ILP, and location potential, we show that our method is more simple and performs better than [7] on the MSRC-21 dataset, 77% to 65% on average recall.

Another system that proposes a simple architecture is [10] which devised a multi-scale Convolutional Neural Network (CNN) to extract features of a pixel for the scene labeling task. Their multi-scale CNN is designed to capture different levels of information, ranging from small region appearance, neighborhood context, and up to the global context of the image. The system remains simple by having only two components: pixel classification and simple post processing to smooth the classification result. However, the system required a specially designed model and careful parameter tuning at training time to get comparable result to the state of the art algorithms. It is therefore hard to re-build the training step and to test on different datasets. They also made a version for RGB-D data from indoor scenes [6]. In our approach, although we use CNN image level feature descriptors, we picked a general purpose feature generator [5] that can be used out of the box without any parameter tuning. At training time, our system needs very little parameter tuning to achieve good results on different datasets.

CRF based Algorithms

CRFs are used in many semantic segmentation algorithms to regularize output labels. The Hierarchical Conditional Random Field (HCRF) [16] uses different levels of quantization, from pixels to segments. They operate under the assumption that there is unlikely to be a single optimal quantization level that is suitable for all object categories.

Beyond regularizing just neighboring pixels, several works model the relationships between all pixels. In the Dense CRF semantic segmentation of [15], mean field approximation and Gaussian filtering is used to make inference in fully connected models practicable. Further, [1] demonstrates that using a Dense CRF to infer all test images at once gives better results than inferring one

image at a time. Our approach shows that a 4-connected neighbor CRF model can achieve results comparable to the fully connected model (both were tested on the MSRC-21 dataset).

More Sophisticated Algorithms

Co-occurrence statistics had been exploited in semantic segmentation systems to boost accuracy. The HCRF was improved further in [17] by incorporating a co-occurrence potential as per-image context information, into the CRF energy function. We propose a simpler system that also uses the co-occurrence statistics, but incorporates them in a different manner. Our system achieves comparable average recall scores to [17], without tuning our parameters per-dataset.

Gonfaus *et al.* [11] proposed another improvement to the HCRF, by adding a new consistency potential to the model, called the harmony potential. The harmony potential encodes all possible combinations of labels, allowing regions to have more than one class, which was a perceived limitation of HCRF models. Further, in [2], they introduced three more cues into the local unary potential to improve recall scores over their previous version. While certainly worthwhile, these algorithms, *e.g.* [2, 11, 16, 17], achieve ever better results by adding complexity to their models. Our system maintains a very plain model, using a simple 4-connected CRF with potts pairwise potentials to encourage harmonization of the neighboring pixels. Our proposed simple system outperforms [11] on both average and global recall scores.

A sophisticated model was successfully demonstrated in [24], where the problems of semantic segmentation, object detection, and scene classification were cast as one holistic CRF model. Their parameters are learned via a structured learning algorithm, and inference is accomplished by a convergent message-passing algorithm. The model exploits various cues, such as scene type, co-occurrence statistics, the shape and location of the object, and different quantization levels to boost the segmentation result. In contrast, our proposed system exploits some of these important cues as context, but integrates them together with a much simpler model, achieving accuracy that approaches that of the more sophisticated model.

Most recently, CNNs have also been exploited in a more sophisticated semantic segmentation framework [13]. They compute feature vectors for each proposed region using two CNN's, trained especially on bounding boxes and free-form versions of the region. Thereafter, the concatenated feature vectors are passed through a linear SVM classifier to get a per-class potential. The final class label for each region is assigned via non-maximum suppression. Although, the system performed very well on the PascalVOC 2012 dataset, it did so at the expense of algorithm complexity.

3. Cheap Semantic Segmentation Model

A good semantic segmentation algorithm should exploit different levels of information: local appearance, global appearance, the context of the scene, and location statistics of objects. We leverage this information with an emphasis on simplicity, so that both large and small training sets can be exploited, and for efficiency at test-time.

3.1. Local Appearance and a Classifier

The cornerstone of visual understanding is having a class-covariant local appearance representation and a compatible classification model. Although superpixels and region-based methods encode neighborhood information, we opt to work on individual raw pixels, curtailing the need to select superpixel algorithms and parameters per dataset.

The Shotton *et al.* STF [22] is one of the simplest useful local appearance classifiers because it processes raw pixel values of a small image patch without constructing a separate feature descriptor. Shotton *et al.* [22] also propose BoST, working in concert with an STF, as a significant contribution, because an STF has limited expressiveness and moderate classification accuracy in itself. The price of BoST’s improved accuracy is its significant computational cost, so we proceed with just the STF model and representation.

3.1.1 Learning from Confusion

The first proposed contribution of this paper is to introduce the Decorrelated Semantic Texton Forest (DSTF), which is an improved version of the STF with only slight additional computational cost at test time. The DSTF emerged from our observation that very similar appearance patches can reach the same leaf node in an STF tree at training-time, even when they have different class labels. This problem occurs in a significant minority of cases. Therefore, the DSTF is designed specifically to reduce the incidence of such high-entropy leaf-nodes.

The DSTF assumes that such “confused” leaf nodes are populated with patches from *distinct* scene types, or categories. To distinguish them, we add an upstream classifier to infer a scene category¹ for each input image. The inferred scene category dictates which single specialist STF should process that image. We train a set of separate STF’s, one for each scene category.

The scene categories are determined automatically, after growing a single temporary STF, depicted at the top of Figure 1. The choice of categories we seek aims to group patches whose visual appearance does not confuse a single STF, and split visually similar patches from different classes that do confuse it.

¹The terms ‘scene category’, ‘scene,’ and ‘cluster’ all refer to the same concept when we are explaining our algorithm.

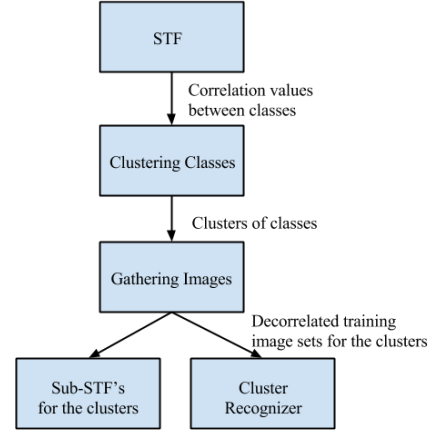


Figure 1. Flow chart for training an upstream scene-category classifier (the Cluster Recognizer), and the downstream Decorrelated Semantic Texton Forest (DSTF) composed of multiple STF’s, with one STF per scene category.

Clustering Classes The temporary single STF is trained with all the images and classes. A label is associated with each pixel, which serves as the center of each small patch. In practice, we re-implement STF’s of the original paper [23] and use the same parameter values, but without including their training invariance, because those parameters were not specified. Our early experiments showed that including some training invariance only improves the STF marginally.

Next, a class correlation matrix Ω is calculated by treating the class distributions at the leaf nodes of the trained STF as observations. Let $\mathbf{X} = \{X_1, \dots, X_T\}$ be the set of class distributions at the T leaf nodes of the entire trained STF. $X_i = \{P(c_1|\mathcal{T}_i), \dots, P(c_C|\mathcal{T}_i)\}$ is a C -dimensional column vector of class probability at the leaf node i , conditioned on training examples \mathcal{T}_i that reached node i . The entries of class correlation matrix Ω are

$$\Omega(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Cov}(x, x) * \text{Cov}(y, y)}}, \quad (1)$$

where $\text{Cov}(x, y)$ is the covariance between class x and y observed from the data \mathbf{X} as row-slices across \mathbf{X} . We then cluster the semantic classes of the original problem by their correlation values, where the distance function is defined by

$$\text{Dist}(x, y) = \Omega(x, y) - \min(\Omega), \quad (2)$$

and $\text{Dist}(x, x) = 0$. We subtract $\min(\Omega)$ from $\Omega(x, y)$ to make the smallest distance equal to zero. We group semantic classes by hierarchical clustering. To commit to hard cluster boundaries, we choose the minimum intra-cluster distance that forces every cluster to have at least three cluster members (classes), to prevent generating trivially

small clusters.

Gathering Images We use the class-clustering to gather the training images into new decorrelated (or less-correlated) training sets. We opt to gather images instead of patches to avoid overfitting. Gathering all images that contain class c would risk piling almost all the training images into some “specialist” STF’s, if a class is prevalent throughout, *e.g.* sky. From experiments on our smallest dataset, MSRC-21’s validation set, we found it already effective to use no more than the top-7% of all training images when training one of the cluster level STF’s.

The procedure to rank the images for each cluster follows. First, the class co-occurrence matrix Ψ is computed from the ground truth images. Each element of Ψ represents the probability of observing a class y given an image of class x , $\Psi(x, y) = P(y|I_x)$. Based on the matrix Ψ , we rank instances (images) for each class c in the cluster separately, by assigning each image the score

$$S(c, G) = \sum_{i \in G} \Psi(c, \mathcal{L}(i)), \quad (3)$$

where i is a pixel in a ground truth image G , and $\mathcal{L}(\cdot)$ returns the ground truth label for the input pixel.

Training Sub-STF’s and the Cluster Recognizer From those decorrelated image training sets, we train separate standard STF’s and the cluster recognizer. The cluster recognizer is a very fast linear SVM, trained on off-the-shelf CNN feature vectors [5]. At test time, the image is fed to the trained cluster categorizer that will redirect the image to an appropriate STF.

A per-class comparison on the MSRC-21 dataset between normal STF and our DSTF is demonstrated in Figure 2. From the figure, we can see that DSTF improves the segmentation accuracies for almost every class. Further analysis is deferred until Section 3.2.1.

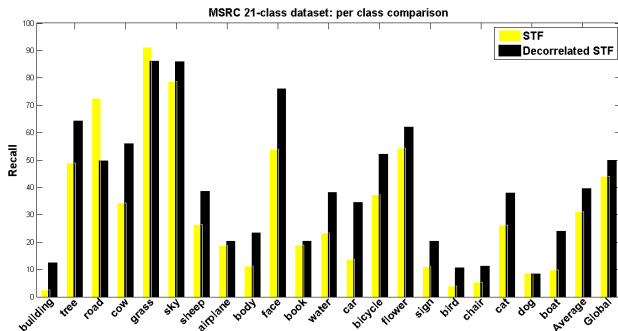


Figure 2. MSRC-21: Per-class comparisons of the DSTF results and STF results, measuring Recall rates.

Efficiency Analysis We compare the efficiency of our proposed DSTF against BoST [22] because BoST is known to

be a real-time semantic segmentation system that can be run on a standard 2008 PC. For simplicity, we count one inference computation of a decision tree or an SVM as one operation. Our goal is to compare the number of operations at test time. For [22], each pixel is routed through another randomized decision forest. At each node, the BoST for a related region of size R pixels is constructed by inferring the STF for every pixel in the region. Hence, the number of operations required for predicting one pixel is $O(lR)$ where l is the number of levels of the randomized decision forest. Whereas the DSTF requires inference by one scene-classifier, a linear SVM, to route the pixel to an appropriate STF, then inference by an STF for that pixel. Therefore, at test time, the DSTF spends merely two operations for one pixel prediction.

3.2. Global Appearance and Objects Co-Occurrence

Experiments from previous work [2, 7, 22, 24] confirm that doing global or local detection concurrently with segmentation can give significantly better segmentation results. In addition, [14, 21] incorporate the co-occurrence statistic as context information, to their detection and/or segmentation system, and showed that such context can improve accuracy. However, to the best of our knowledge, there is no work that trains using the two cues of a) the presence of the object in the image and b) such context information, together, to improve the segmentation process.

Our second contribution is to propose the Context Sensitive Image Level Prior (Context Sensitive ILP or ILPcont). Please note that, the terms context and co-occurrence will be used interchangeably from now on. From the experiment, Table 1 shows that selecting the algorithm, which is aware of the co-occurrence of classes, to generate the Image Level Prior (ILP) produces much more promising results than the algorithm that trained to detect each class separately. The details of the Context Sensitive ILP are discussed next.

3.2.1 Context Sensitive Image Level Prior

Although Image Level Prior or image level class detection has been proved to be useful for semantic segmentation in recent papers, *e.g.* [2, 22], previously the ILP only models the presence of classes for a single image. We propose a new image level class detection that takes into account co-occurrence statistics of the classes in the entire training data. The use of multi-label randomized trees allows us to model the global appearance of a single image and the co-occurrence statistics of the classes of entire training data at the same time. The multi-label randomized trees is first proposed in [8], but it was used in different paradigm which is predicting a structured output. We, contrastingly, use the al-

gorithm to learn both implicit class co-occurrence statistics of the training data and presences of classes in an image. Even though the algorithm is used to approach slightly different problems, the algorithm can be directly applied without any modification.

In this section, we will give a brief explanation of the multi-label randomized trees algorithm [8]. Multi-label random forests are random forests with a minor modification on the splitting quality metric. The metric is used to measure how well a feature splits the data at the node. The modification is made to take into account more than one class for the node splitting instead of a single class in the original randomized tree based model. The metric is based on the Gini entropy, and the modification are as follows,

$$\text{score}(\mathcal{T}, \mathcal{F}) = \frac{1}{C} \sum_{k=1}^C \text{score}^k(\mathcal{T}, \mathcal{F}), \quad (4)$$

$$\text{score}^k(\mathcal{T}, \mathcal{F}) = G_k(\mathcal{T}) - G_{k|\mathcal{F}}(\mathcal{T}), \quad (5)$$

$$G_k(\mathcal{T}) = 2 \left(\frac{\sum_{t_i \in \mathcal{T}} \mathcal{L}_k(t_i)}{n} \left(1 - \frac{\sum_{t_i \in \mathcal{T}} \mathcal{L}_k(t_i)}{n} \right) \right), \quad (6)$$

$$G_{k|\mathcal{F}}(\mathcal{T}) = G_k(\mathcal{T}_l) + G_k(\mathcal{T}_r), \quad (7)$$

where \mathcal{T} is data of size n that reaches the node and \mathcal{F} denotes a test function that routes subset of the data \mathcal{T}_l to its left child node when all member of \mathcal{T}_l satisfy \mathcal{F} otherwise routes the data \mathcal{T}_r to the right child node. C is the number of semantic classes. $\mathcal{L}_k(t_i)$ is the function that returns 1 when $L_{ik} = 1$, and 0 otherwise; and L_{ik} indicates the presence of class k in the datapoint i^{th} .

Table 1 compares results of the original Semantic Texton Forests with our 2 proposed components. Please note that, combining the original STF with context sensitive ILP outperforms the full model of BoST that proposed in [22]; our system has average recall 68.03% compared to 66.9% of [22].

Method	no ILP	normal ILP	context ILP
STF (average)	31.02%	35.29%*	68.03%
DSTF (average)	39.67%	41.56%	70.07% †
STF (global)	44.00%	50.16%*	72.21%
DSTF (global)	49.82%	55.52%	73.97% †

Table 1. MSRC-21: Average and Global recalls of Decorrelated Semantic Texton Forests and Context Sensitive ILP (†) and original Semantic Texton Forests and ILP (*).

3.3. Location Potentials

The last crucial ingredient are the location potentials. The location potentials are simply the statistics of how likely each absolute location in the image to be occupied by particular classes. The location potential is also used in [23].

In this work, training images are first split into 2 groups: portrait images and landscape images. Next, for each group, we count the frequencies of each absolute location to be landed by a particular class. After having the location potentials for each class (each class has 2 location potentials: portrait and landscape), the location potentials are used as look up tables for an input location.

Figure 3 illustrates the importance of the location potentials comparing to DSTF.

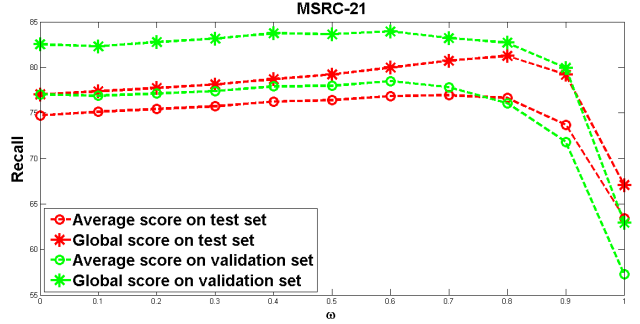


Figure 3. MSRC-21: Influence of location potentials, ranged from using pure DSTF, $\omega = 0$, to pure location potentials, $\omega = 1$ in the model. Note that, we optimise the system on average per-class recall.

3.4. Integration of the Components

A simple Conditional Random Field model is selected to assemble the components. Simple, in our case, refers to utilising an ordinary grid graph, where each pixel has 4-connected neighbors, and the potts pairwise potential. This potential assigns low energy to two adjacent pixels with the same class, and high energy otherwise.

More formally, We cast the inference of our system as a minimization of the energy function

$$E(x_1, \dots, x_N | I) = \sum_{i \in N} \zeta_I \phi_i(x_i) + \sum_{(i,j) \in P} \psi_{i,j}(x_i, x_j), \quad (8)$$

where x_i is a random variable associated with the test image I . The random variable x_i can be assigned one of the labels in the label set $\mathbf{c} = \{c_1, \dots, c_C\}$. N is the number of pixels in the test image $I = \{x_1, \dots, x_N\}$, and P is a set of all pairs of neighboring pixels. ψ is defined as the potts potential and ϕ_i is the unary potential defined as,

$$\phi_i(x_i) = (1 - \omega) * DSTF(x_i) + \omega * Location(x_i), \quad (9)$$

and ζ_I is the image level prior of image I . Minimization is carried out using the graph cuts algorithm of Boykov *et al.* [3].

Figure 6 compares the average per-class recall results of our whole system, and when certain components are missing. The blue dotted line shows our results when all components were trained with standard data splitting, as per [23].

The red dotted line shows average per-class recalls of the system when only the ILP was trained with additional data, sampled from the test data, so the test data size is smaller than the standard one. The yellow dotted line shows the result when the ILP was trained in the unusual ways: the black star is the result when ILP was trained on the entire dataset, with no unseen data for the ILP, and the magenta star represents the result of Ideal ILP, assuming that we know the actual image tags.

4. Results by Dataset

We evaluate our approach via two well-known semantic segmentation datasets, MSRC-21 and PascalVOC-2010. MSRC-21 is now considered a small older dataset, but it is commonly used for validating semantic segmentation algorithms. Whereas the latter, PascalVOC-2010, is one of the newest and largest datasets. We choose these datasets to prove that our approach is robust with limited training data, as well as with great diversity of scenes. The main reason we select PascalVOC-2010 over the newer or the older versions of the same competition is the recently published finer ground associated with it [20].

4.1. MSRC-21 [23]

The MSRC-21 dataset is composed of 591 images of size 320 x 213 and 213 x 320. The segmentation ground truth is made up of 21 classes which are mixed between background classes and object classes. The parameters we are using for the MSRC-21 dataset are the same as in [22], with one additional parameter ω to weight between the appearance potential (DSTF) and the location potential. We tune the extra parameter using the validation set as shown in Figure 3.

Table 1 demonstrates that both the proposed DSTF and context sensitive ILP work to complement each other, and give $\approx 9\%$ and $\approx 37\%$ improvement respectively. Figure 6 shows that our full model (integrating DSTF, Context Sensitive ILP, and Location potential by simple CRF) can achieve a result that is comparable to state of the art results. Besides, when the ILP has more training data, our proposed model even beats the top algorithm of this dataset. Table 4 shows detailed results for each class compared to state of the art algorithms. One can observe that our system performs well on all classes. Qualitative results can be found in Fig. 4.

4.2. PascalVOC-2010 [9]

The PascalVOC-2010 semantic segmentation dataset consists of 964 training images, 964 validation images, and 964 test images. The dataset has 20 object classes and 1 background class which includes everything but the 20 object classes. The background class occupies 60.1% of all pixels in the training and validation set [20]. As the ground

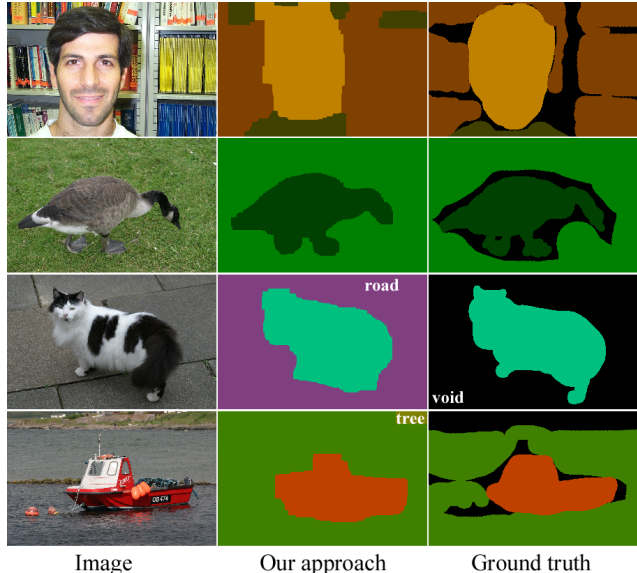


Figure 4. Qualitative results on the MSRC dataset.

truth for the test set is not publicly available, the organizers run an evaluation server where users can submit up to two submissions per week. We test our algorithm on this dataset using default parameters, *i.e.* the same parameters used for the MSRC-21 dataset. In addition, [20] relabeled the ground truth by adding more classes and removing the background class. We also show the results of training our ILP with this new ground truth data on the standard 20-object-classes, and with additional context classes, such as water, sky, road, etc. (+Add_context).

Table 2 illustrates our quantitative results on the test data. Alone, STF and DSTF do not perform well with this dataset, since the data is more diverse and has a very large and complex background class. However, the Context Sensitive ILP still produces impressive improvements, improving the result by $\approx 17\%$ over the pure STF and DSTF, comparing to an improvement of only $\approx 7\%$ by using the multi-class ILP. Interestingly, our Context Sensitive ILP coupled with just Location potentials is proving very powerful, despite missing out on substantial information available to the full system.

Although the new ground truth on VOC has better ground truth, we can see that the accuracy decreases. The relabeling process has modified the ground truth for the standard 20-object-classes, but we still evaluate the result via the evaluation server which evaluates based on the old ground truth. Furthermore, adding more context classes can hurt the accuracy of the system because a larger number of classes reduces the ILP prediction accuracy. Fig. 5 demonstrates some qualitative results.

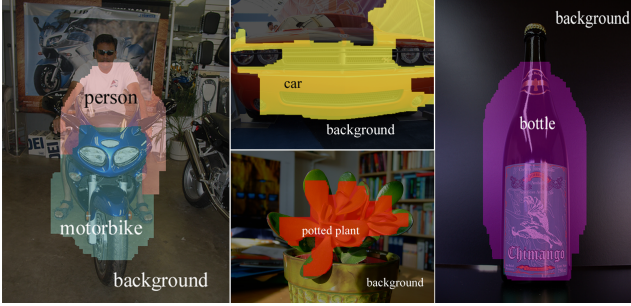


Figure 5. Qualitative results on the PascalVOC-2010 dataset when using our approach. We illustrate the results overlaid on the image since the ground truth for is not available.

Methods	IoU
<u>Train on original ground truth [9]</u>	
STF	1.454
DSTF	0.858
STF+ILPmult	6.019 \ddagger
STF+ILPcont	16.656 \ddagger
DSTF+ILPcont	16.947 \ddagger
Location+ILPcont	21.474 \star
STF+Loc+ILPmult	7.060 \ddagger
STF+Loc+ILPcont	21.403 \ddagger
Our full model (DSTF+Loc+ILPcont)	21.588 \ddagger
Our full model (DSTF+Loc+ILPcont)	24.058 \star
<u>Train ILP on the new ground truth [20]</u>	
STF+Loc+ILPmult	6.585 \ddagger
STF+Loc+ILPcont	17.760 \ddagger
STF+Loc+ILPcont+Add_context	16.875 \ddagger
<u>State of the art algorithms</u>	
Topic model [12]	27.8
DenseCRF [15] (non-standard test set)	30.2
HCRF+Cooc [17]	30.3
Whole [24] (non-standard test set)	31.2
Harmony ₄ [2]	38.0
Composite [19]	49.6

Table 2. Intersection over Union score of the system on PascalVOC-2010 dataset. We demonstrate the results from different combinations of our proposed components. Please note that, ILPmult and ILPcont represent the multi-class image level prior and context sensitive image level prior respectively. Loc stands for location potential, and Add_context represents us training the ILP with extra context classes: sky, road, building, water, grass. \ddagger indicates that the methods use that same set of parameters, to make the numbers comparable; \star indicates the parameter was tuned by cross validating, fixing the ILP for class background to probability 0.1, 0.2, ..., 1.0.

5. Limitations

We validated our proposed system on another standard dataset, that demonstrates a predictable limitation of our

approach. The CamVid dataset [4] consists of image sequences of road scenes, where the ground truth labels associate each pixel with one of the grouped 11 semantic classes. To be comparable to other algorithms, we down-sample all the images by a factor of 3 as in [18].

Table 3 shows the quantitative results of our algorithm on the CamVid dataset. Since most of the images have almost the same set of classes present in them, context information is not useful here. Therefore, we can see that the Context Sensitive ILP performed worse than the normal multi-class ILP because the context sensitive ILP can extract very few co-occurrence patterns from the training data. DSTF also hurts accuracy because the training sub-STFs are not really specialized, *i.e.* they have access to artificially small training sets, with little difference between scene categories.

Methods	Average	Global
STF	29.95	27.25
STF+ILPcont	27.31	27.20
STF+ILPmult	29.53	29.38
DSTF	10.41	7.38
DSTF+ILPmult	26.84	28.32
Loc+ILPmult	27.85	55.96
STF+Loc+ILPmult	40.27	59.39
Full model (DSTF+Loc+ILPmult)	34.56	52.23
<u>State of the art algorithm</u>		
Combining Object Detection [18]	62.5	83.8

Table 3. Average and Global recalls of the system on CamVid dataset. We tested our proposed model on different combinations of the components.

6. Conclusion

We have shown that a combination of simple techniques can yield excellent accuracy, given only a limited computational budget. Our DSTF shows an impressive ability to empower the inaccurate appearance predictions of a normal STF, with only a small extra overhead. This is noteworthy because each sub-STF is working with less training data. The Context Sensitive ILP proved quite capable of recovering from even fairly bad appearance predictions. While other ILP models have been proposed previously, using the co-occurrence statistics jointly with image level class detection can now be accomplished cheaply, and can yield a substantial improvement in accuracy.

We are making our code publicly available. Many extensions for the future are possible because the existing system is simple and complementary to many other approaches. A natural extension would use fast filters over the image as extra appearance channels in the DSTF. It could also be fruitful to learn a variety of location potentials, *i.e.* for different camera poses, *e.g.* from car-mounted or hand-held cameras.

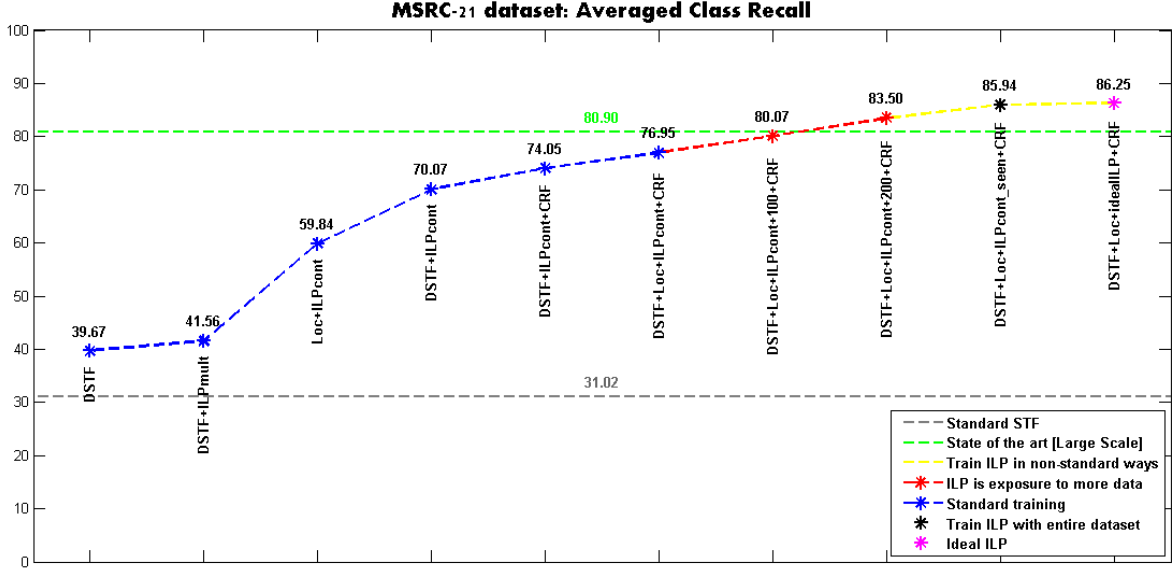


Figure 6. Average per-class recall results of the system (best viewed in colors) with or without certain components. The results are compared to the state of the art [1], green dotted line. Important notations: ILPmult (normal multi-class ILP regardless of the co-occurrence statistic), ILPcont (the proposed Context Sensitive ILP), ILPcont+N (the proposed Context Sensitive ILP when training with standard training and validation data + N images sampled from standard test data and testing on unseen data, therefore test data - N sampled images), ILPcont_seen (the context sensitive ILP trained on all data; thus no unseen data for the ILP), and idealILP (When the ILP component produce 100% correct prediction), and Loc (Location potentials).

	building	tree	road	cow	grass	sky	sheep	aeroplane	body	face	book	water	car	bicycle	flower	sign	bird	chair	cat	dog	boat	Average	Global
BoST [22]	49	79	78	97	88	78	97	82	66	87	93	54	74	72	74	36	24	51	75	35	18	67	72
Harmony ₁ [11]	60	77	76	91	78	88	68	87	56	73	95	76	77	93	97	73	57	81	81	46	46	75	77
HCRF+Cooc [17]	82	88	93	73	95	100	88	83	65	88	85	92	87	88	96	96	27	37	49	80	20	77	87
DenseCRF [15]	75	91	90	84	99	95	82	82	80	89	98	71	90	94	95	77	48	61	78	48	22	78	86
Whole [24]	71	90	89	79	98	93	86	88	68	90	97	86	84	94	98	76	53	71	83	55	17	79	86
Harmony ₄ [2]	66	84	82	81	87	93	83	81	70	78	90	82	86	94	96	87	48	81	82	75	52	80	83
Large Scale [1]	73	90	90	85	99	95	82	86	87	91	96	74	88	91	96	83	54	79	81	60	18	81	87
Our	51	79	85	92	96	81	90	68	59	93	95	84	76	92	98	75	64	71	86	47	34	77	80
Our+100	58	86	89	92	97	87	93	67	61	90	94	87	75	92	100	88	70	83	77	91	70	83	85
Our+200	71	86	88	94	98	86	95	79	68	83	99	86	76	96	100	100	91	94	80	85	51	86	88
Our+Seen	68	88	87	94	97	89	94	71	60	92	99	92	85	93	98	93	78	96	89	81	61	86	88
Our+Ideal	71	88	87	94	96	90	95	70	59	91	99	91	83	89	98	92	79	97	90	81	71	86	88
Harmony ₄ +Ideal [2]	68	92	89	86	93	97	88	91	60	73	100	85	86	94	100	89	77	96	95	94	74	87	89

Table 4. MSRC-21 segmentation results. Note that we show results of our system and [2] with Seen ILP and Ideal ILP to show the upper bound of the systems thus we do not include them to the comparison.

References

- [1] J. M. Alvarez, M. Salzmann, and N. Barnes. Large-scale semantic co-labeling of image sets. In *IEEE Winter Conference on Applications of Computer Vision*, pages 501–508, Mar. 2014.
- [2] X. Boix, J. M. Gonfaus, J. van de Weijer, A. D. Bagdanov, J. Serrat, and J. González. Harmony Potentials Fusing Global and Local Scale for Semantic Image Segmentation. *International Journal of Computer Vision*, 96(1):83–102, Apr. 2011.
- [3] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *Pattern Analysis and Machine Intelligence 2001*, (November):1222–1239, 2001.
- [4] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and Recognition using Structure from Motion Point Clouds. In *ECCV*, pages 1–14, 2008.
- [5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the Devil in the Details : Delving Deep into Convolutional Nets. In *BMVC*, 2014.
- [6] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor Semantic Segmentation using depth information. In *International Conference on Learning Representation*, pages 1–7, 2013.
- [7] G. Csurka and F. Perronnin. A Simple High Performance Approach to Semantic Segmentation. In *BMVC*, pages 22.1–

- 22.10. British Machine Vision Association, 2008.
- [8] M. Dumont, R. Maree, L. Wehenkel, and P. Geurts. Fast Multi-class Image Annotation with Random Subwindows and Multiple Output Randomized Trees. In *VISAPP*, 2009.
 - [9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. <http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html>.
 - [10] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning Hierarchical Features for Scene Labeling. *PAMI*, pages 1–15, 2013.
 - [11] J. M. Gonfaus, X. Boix, J. van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez. Harmony Potentials for Joint Classification and Segmentation. *CVPR*, pages 3280–3287, June 2010.
 - [12] I. González-Díaz and F. Díaz-de María. A region-centered topic model for object discovery and category-based image segmentation. *Pattern Recognition*, 46(9):2437–2449, Sept. 2013.
 - [13] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous Detection and Segmentation. In *ECCV*, pages 1–16, 2014.
 - [14] P. Kotschieder, S. Rota Buló, A. Criminisi, P. Kohli, M. Pelillo, and H. Bischof. Context-Sensitive Decision Forests for Object Detection. In *NIPS*, pages 1–9, 2012.
 - [15] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 1–9, 2011.
 - [16] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Associative Hierarchical CRFs for Object Class Image Segmentation. In *ICCV*, 2009.
 - [17] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr. Graph Cut Based Inference with Co-occurrence Statistics. In *ECCV*, pages 1–14, 2010.
 - [18] L. Ladicky, P. Sturges, K. Alahari, C. Russell, and P. H. S. Torr. What , Where & How Many ? Combining Object Detectors and CRFs. In *ECCV*, 2010.
 - [19] F. Li, J. Carreira, G. Lebanon, and C. Sminchisescu. Composite Statistical Inference for Semantic Segmentation. In *CVPR*, number 1, pages 3302–3309. Ieee, June 2013.
 - [20] R. Mottaghi, X. Chen, X. Liu, N. Cho, S. Lee, S. Fidler, R. Urtasun, and A. Yuille. The Role of Context for Object Detection and Semantic Segmentation in the Wild. In *CVPR*, 2014.
 - [21] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in Context. In *ICCV*, pages 1–8. Ieee, 2007.
 - [22] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, pages 1–8, June 2008.
 - [23] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Texton-Boost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *International Journal of Computer Vision*, 81(1):2–23, Dec. 2007.
 - [24] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, pages 702–709, June 2012.